Synthetic Dataset Report

Original Dataset Overview:

The original dataset, Coles_cleaned.csv, contains **20,608 rows** and **8 columns**, representing product-level data from a retail store. Key fields include:

- product_code: Unique identifier
- category: Product category
- item_name: Name of the item
- best_price and item_price: Pricing metrics
- unit_price and best_unit_price: Unit-based pricing
- link: Product URL

This dataset formed the basis for creating a synthetic dataset simulating weekly discounts over an 8-week promotional period.

Synthetic Dataset Summary:

The cleaned synthetic dataset, Coles_synthetic_8weeks_v3_cleaned.csv, was derived using a custom rule-based discount logic. It contains:

- **164,864 rows**
- **19,782 unique products**
- **8 weeks of coverage**
- Each product appears **exactly once per week**

Discount Strategy Logic

1. One **random brand per category per week** receives a **50% discount** (excluding Coles).
2. **30% of remaining brands** get randomly assigned **20% or 30% discounts**.
3. **20% of remaining brands** get **10% discounts**.
4. **Coles** brand items only receive **20% or 30% discounts**, never 50%.
5. All other items remain **at full price**.
6. All **discounted prices** were **rounded up to the nearest $0.50**.

Dataset Quality Assessment:

Quality Assurance Checks:

Performed using the script quality_check.py and notebook advanced_quality_check.ipynb.

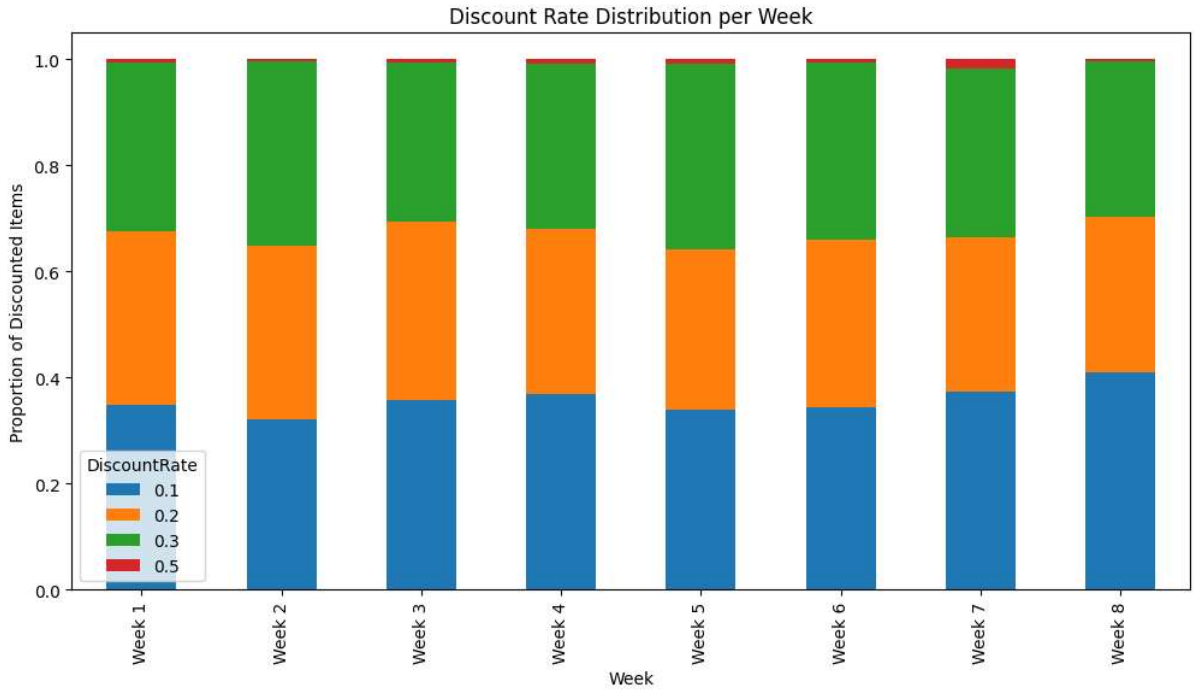
Category	Description	Result
Missing Values	No missing values across columns	✅ Pass
Weekly Coverage	All 8 weeks present, each product appears once per week	✅ Pass
Discount Logic	Coles never received 50% off, only one 50%-off brand per week	✅ Pass
Discount Accuracy	Discounted prices correctly rounded to $0.50	✅ Pass
Price Outliers	Prices capped to a min of $1.00 and max of $100.00	✅ Pass
Clustering Check	4 clean clusters detected via K-Means	✅ Pass
Z-Score Outliers	1.8% rows flagged, mostly minor edge cases	✅ Pass

Visualizations and Interpretations:

- **K-Means Cluster Distribution:**

  - **What it shows:**
    The results of clustering all rows based on price and discount-related features.
  - **Interpretation:**
    - Cluster 0: Regular-priced or non-discounted items
    - Cluster 2 & 3: Discounted but moderately priced items
    - Cluster 1: High-end or rare items, either expensive or deeply discounted
  - **Significance:**
    - The presence of **4 distinct clusters** indicates the dataset has **good structural segmentation**.
    - This ensures LightGBM can **differentiate between types of promotional behaviors** during training.
    - Lack of micro-clusters or noise post-cleaning confirms **data quality and logical grouping**.

- **Boxplots of Discounted Prices (Weeks 1, 4, 8)**



- **What they show:**
  Category-wise spread of DiscountedPrice values for select weeks. These visualizations check for:
  - Outliers
  - Central tendency (median)
  - Price spread consistency
- **Interpretation:**
  - Median and quartile ranges are **stable across weeks**, showing **price consistency**
  - No values exceed $100 or fall below $1, confirming effective **price capping**
  - No category dominates or exhibits extreme variance, implying **pricing fairness** across product types
  - Supports model training by providing **stable learning boundaries** without skewed data